EFFECT SIZE REPORTING IN CURRENT HEALTH EDUCATION LITERATURE

Matthew Lee Smith, MPH, CHES, CPP
Heather H. Honoré, PhD, MSW

Abstract: This study evaluated four health education journals for prevalence of effect size reporting among articles using statistical analysis techniques yielding variance-accounted-for effect sizes (i.e., analyses of variance and regression). One hundred thirty-six articles, meeting exclusionary criteria, were analyzed for effect size reporting using a content analysis rubric. Of the articles examined, 133 articles (97.79%) reported statistical significance test results (i.e., p values); however, only 43 reported variance-accounted-for effect sizes (31.62%) and 11 (8.09%) interpreted these effect sizes in terms of practical significance. Based on study results, effect sizes and corresponding practical significance should be uniformly reported in health education literature.

INTRODUCTION

The variety of ways to report statistical results has contributed to debate regarding appropriateness of statistical significance testing in health education research. Several researchers recommend reporting effect size in conjunction with statistical significance testing (Anderson & Burnham, 2000; Cohen, 1994; Fidler, Geoff, Mark, & Neil, 2004; Hyde, 2001; Thompson, 1999a, 1999b; Wilkinson, 1999). Despite health education researchers' frequent use of statistical analyses yielding effect sizes, adherence to such recommendations is not uniformly represented in the literature. Currently, effect size reporting is not required in publication submission guidelines. Yet, clearly understanding the merit and purpose of reporting effect size may encourage health education researchers to report effect sizes.

COMMONLY USED STATISTICAL ANALYSES IN HEALTH EDUCATION RESEARCH

With over sixty-one-identified effect sizes (Elmore & Rotou, 2001), researchers have a myriad of opportunities to interpret and report research findings. Thompson (2006) states that "Effect size is a statistic quantifying the extent to which sample statistics diverge from the null hypothesis" (p. 187). Effect sizes are classified into three major classes, standardized differences effect sizes, variance-accounted-for effect sizes (VES), and "corrected" effect sizes (Vacha-Haase & Thompson, 2004). (See Cohen, 1977; Hojat & Xu, 2004; Watkins, Rivers, Rowell, Green, & Rivers, 2006; or Zakzanis, 2001 for additional details concerning effect size calculations.)

Analyses of variance (i.e., ANOVA, MANOVA, ANCOVA) and multiple regression analyses persist as frequently utilized statistical methods in social science research (Goodwin & Goodwin, 1985; Kirk, 1996). Based on the general linear model (GLM), these similar statistical techniques yield VES (Thompson, 2006). Variance-accounted-for effect sizes have scores ranging from zero (0) to one (1) and represent the total percentage of variance explained by independent (predictor) variables (Tapia & Marsh, 2002). ANOVA generates the effect size eta^2 (η^2 or correlation ratio), which is computed by dividing the sum of squares explained (also termed between, model, and regression) by the sum of squares total (Thompson, 2002a). Whereas, multiple regression generates the effect size multiple r squared (R^2), which is also computed by dividing the sum of squares explained by the sum of squares total (Thompson, 2002a).

PRACTICAL IMPORTANCE OF EFFECT SIZE REPORTING

Statistical significance testing allows researchers to examine the probability of sample results based

Matthew Lee Smith, MPH, CHES, CPP, is affiliated with Child & Adolescent Health Research Lab, Department of Health & Kinesiology, Texas A&M University. Heather H. Honoré, PhD, MSW, is a VA HSR&D Post-Doctoral Fellow in the Michael E. DeBakey VA Medical Center, Houston Center for Quality of Care & Utilization Studies (HCQCUS). Please address all correspondence to Matthew Lee Smith, MPH, CHES, CPP, Child & Adolescent Health Research Lab, Department of Health & Kinesiology, Texas A&M University, 4243 TAMU, College Station, TX 77842. Telephone: 979.845.3290. Fax: 979.847.8987. matlsmit@tamu.edu.

on a set of assumptions. Despite conventional acceptance, this type of testing lacks the ability to address the relative importance of research findings (Thompson, 2003). Calculating effect size not only identifies differences between hypotheses (p values) but also the size and associated power of such differences. This enables researchers to formulate conclusions regarding context-specific practical significance (Harris, 1991; Kirk, 1996). Examining effect sizes across studies enables comparison of several key domains, including internal experimental validity (i.e., addressing the "tightness" of the study) and external experimental validity (i.e., addressing generalizability of the study to similar studies or demographics; Elmore & Rotou, 2001).

EFFECT SIZE REPORTING IN HEALTH EDUCATION LITERATURE

No requirements for reporting effect sizes exist among health education scholarly journals, despite 1994 and 1999 reporting recommendations and "encouragement" to supplement p values with effect sizes (American Psychological Association, 1994; Watkins et al., 2006; Wilkinson, 1999). In addition to recommendations to supplement statistical significance testing in health education literature (Buhi, 2005), three specific reporting guidelines have been offered in accordance with the Task Force on Statistical Inference (Thompson, 2002b). These guidelines are to: 1) report effect sizes, 2) report confidence intervals, and 3) use graphics.

Although "encouragement" to report effect sizes along with p values is thought to increase effect size reporting (Kirk, 2001; Nix & Barnette, 1998), authors remain unconvinced that mere "encouragement" has been sufficient (Kirk, 1996; McLean & Ernest, 1998; Thompson, 1998; Vacha-Haase & Thompson, 1998). Moreover, researchers report infrequent compliance with the aforementioned recommendations (Keselman & Huberty, 1998).

The purpose of this article is to examine current literature in predominant health education journals regarding adherence to recommendations for reporting effect sizes in order to compliment and/or supplement current practices of statistical significance test reporting. Specifically, the aims were as follows: 1) identify the number of recently published health education articles utilizing analyses of variance and/or regression analyses (i.e., that have the ability to yield effect sizes); 2) identify the number of recently published health education articles utilizing analyses of variance and/or regression analyses that also reported effect sizes; and 3) identify the number of recently published health education articles utilizing analyses of variance and/or regression analyses that reported effect sizes and provided explanation of the practical significance of those effect sizes.

METHODS

A content analysis was conducted using articles published within four health education journals over the previous two years. This analysis evaluated adherence to effect size reporting recommendations. All 2006 and 2007 issues of Health Education Research (HER), Health Education and Behavior (HEB), Health Education Journal (HEJ), and International Electronic Journal of Health Education (IEJHE) were analyzed. These journals were selected based on their submission guidelines requiring that articles submitted for publication contain only health education-related content. None of the four journals selected mandate effect size reporting within the author submission guidelines

INCLUSION AND EXCLUSION CRITERIA

Selected journal articles contained primary data collection and were empirically driven. Selected articles were included only if methods of statistical analysis included analyses of variance (i.e., ANOVA, MANOVA, ANCOVA) and/or regression analyses. Omitted journal articles included secondary analyses, book reviews, commentaries, editorials, presidential addresses, award papers, perspectives, conceptual articles, systematic/literature reviews, policy reviews, non-empirical qualitative research, process evaluations, instructions for authors, reference indices, erratum, and statistical analyses not encompassing analyses of variance or regression analyses. Articles within each journal were omitted based on exclusionary criteria.

INSTRUMENT

Journal articles were evaluated utilizing a content rubric. The content rubric contained analysis criteria including sample size(s), study design, study methodology and research focus. Additionally, the rubric contained criteria related to effect sizes and types, p values, types of statistical analyses reported, and whether or not effect sizes were discussed in the results or discussion sections of the article. For the purposes of this article, only criteria related to statistical analyses and effect size reporting were included.

PROTOCOL

Methods, results, and discussion sections of journal articles were reviewed for content regarding effect size reporting. Predetermined inclusion/exclusion criteria were followed to omit non-relevant journal articles. Information obtained from each article meeting inclusionary criteria was recorded in the content rubric. Rubric cells were marked with an "x" to indicate the article included the analysis criteria of effect size reporting, p value reporting, and

whether or not effect sizes and their practical significance were discussed within the article. Categories including p-values, confidence intervals, odds ratios, VES, and discussion of effect size's practical significance were recorded in the content rubric for each corresponding article. Column totals were summed and frequencies were calculated.

RESULTS

SAMPLE

After reviewing current issues of HER, HEB, HEJ, and IEJHE, a total of 383 articles were examined for potential inclusion (156, 128, 68, and 31, respectively). Upon application of exclusionary criteria, 122 articles were omitted (37, 54, 26, and 5, respectively). The remaining 261 articles were examined for content regarding effect size reporting (119, 74, 42, and 26, respectively). After reviewing articles for statistical analyses potentially yielding VES, 136 articles were included within the final sample (see Tables 1 and 2).

TYPES OF STATISTICAL ANALYSES IN QUALIFYING ARTICLES

Of the articles meeting predetermined inclusion/exclusion criteria, various types of statistical analyses were reported. In some cases, more than one statistical technique was utilized. Of the 158 statistical analyses (n=136 articles) in qualifying articles that also yield VES, the majority used regression (61.39%). Analysis of variance (33.54%), path analyses (1.27%), structural equation modeling (3.17%), and canonical correlations (0.63%) were also present (see Table 3).

Table 1. Frequency of Articles Meeting Criteria by Year

	Year	Overall Content	Primary Data Articles	Articles Meeting Criteria	Primary Articles Meeting Criteria (%)
All Journals					
	2006	191	121	69	57.02
	2007	192	140	67	47.86
	Total	383	261	136	52.11

Table 2. Frequency of Articles Meeting Criteria by Journal

	Year	Overall Content	Primary Data Articles	Articles Meeting Criteria	Primary Articles Meeting Criteria (%)
HER					
	2006	73	51	34	66.67
	2007	83	68	39	57.35
	Total	156	119	73	61.34
HEB					
	2006	65	34	19	55.88
	2007	63	40	22	55.00
	Total	128	74	41	55.41
HEJ					
	2006	34	20	8	40.00
	2007	34	22	4	18.18
	Total	68	42	12	28.57
IEJHE					
	2006	19	16	8	50.00
	2007	12	10	2	20.00
	Total	31	26	10	38.46

Table 3. Statistical Analyses Used in Qualifying Articles (n=158)

	REG (%)	ANOVA (%)	PATH (%)	SEM (%)	CANONICAL (%)
HER (n=86)	62.79	31.40	2.33	3.49	0.00
HEB (n=48)	62.50	33.33	0.00	4.17	0.00
HEJ (n=14)	42.86	50.00	0.00	0.00	7.14
IEJHE (n=10)	70.00	30.00	0.00	0.00	0.00

Table 4. Effect Size and Significance Reporting in Qualifying Articles (n=136)

	P value (%)	CI/OR (%)	Effect Size (%)	Practical Significance Discussion (%)
HER (n=73)	97.26	50.68	36.99	5.48
HEB (n=41)	100.00	41.46	31.71	14.63
HEJ (n=12)	91.67	33.33	8.33	0.00
IEJHE (n=10)	100.00	70.00	20.00	10.00

EFFECT SIZE REPORTING IN QUALIFYING ARTICLES

There were136 articles within the final sample. Of these articles, 133 articles (97.79%) reported statistical significance test results (i.e., p values). However, only 43 articles reported VES (i.e., $\eta 2$ and R^2; 31.62%) and 11 (8.09%) interpreted these effect sizes in terms of practical significance. Fifty-five articles reported confidence intervals and/or odds ratios (40.44%; see Table 4).

DISCUSSION

Three main objectives have been addressed in this article. The first objective was to highlight benefits of incorporating variance-accounted-for effect size reporting within the health education literature. Despite common practices of reporting statistical significance testing, acceptance of p values as measures of practical significance or replicability may lead to incorrect interpretation of results in health education research (Cohen, 1977, 1990, 1994; Kirk, 1996, 2001): Reporting and interpreting only p values may also lead to erroneous recommendations for practice or future research. Further, misinterpretation of results due to using p values as measures of practical significance may foster poor statistical reporting practices across disciplines.

The second objective was to support recommendations for reporting effect sizes in health education literature. These recommendations have been endorsed by leading social science researchers including Thompson, Cohen, Kline, and Vacha-Haase. Emergence of recommendations by such prominent figures suggests a requisite paradigm shift in statistical analyses reporting styles. These recommendations should be adopted by health education researchers.

To increase effect size reporting, journal editors, researchers, and authors of textbooks and publication manuals should unite views regarding the importance of effect size reporting and the practical limitations of statistical significance testing (Kirk, 1996; Vacha-Haase, 2001). A common vision of statistical analyses reporting will promote uniformity of materials submitted for publication, thus allowing researchers to better interpret study findings (i.e., understand the size of hypotheses differences, rather than merely the existence of differences); make reasonable comparisons between studies in related fields; and form inferences about replicability. A common style for reporting statistical analyses may facilitate success related to individual and community behavioral change, elimination of health disparities, and improved quality of life for all citizens.

Based on the results of this study, the authors recommend that health education journal editors enforce strict author submission guidelines for reporting effect sizes. These guidelines should be initiated and supported by authors and publishers (Buhi, 2005; Kirk, 1996; Thompson, 2006; Vacha-Haase, 2001; Watkins et al., 2006). The following examples were selected from qualifying articles to promote uniform reporting and interpretation within the health education literature. These heuristic examples explicitly report effect sizes and interpretations of these effect sizes in terms of practical significance:

> Despite statistically significant change across all outcome variables, the effect sizes were small. What is the practical importance of these results? Three of the outcome variables (i.e., carrying medication list, reading labels, and talking with health care provider) were concerned with preventing infrequent but serious events (e.g., adverse side effects or drug-dietary supplement interactions). For example, the number of participants reporting that they spoke with their health

> care provider about supplement use increased by approximately 12% in the experimental group compared with 3% in the comparison group. (Mitchell, Ash, & McClelland, 2006, p. 387)
>
> The proportion of variance in intention explained by the three variables (attitude, subjective norms, perceived behavioral control) was 27%, which is consistent with the results of two similar studies by Rutter (2000) and Drossaert et al. (2003), which explained 29% and 48.6% of the variance, respectively. The low percentage of variance explained in this study can be attributed to the fact that breast cancer screening is a complex behavior, which entails personal, social, and environmental factors in its explanation. (Tolma, Reininger, Evans, & Ureda, 2006, p. 244)

The authors recommend that these heuristic examples be used as guides for effect size reporting and interpretation when submitting manuscripts for publication. (More detailed examples are available in Hojat & Xu, 2004.)

The third objective of this article was to examine current literature to identify VES reporting. Findings from this content analysis indicate that analyses of variance and regression analyses are common within primary data-focused articles in health education literature (see Table 3). Yet, 64.49% of the articles published in the health education literature remain either descriptive, conceptual, or utilize statistical analysis techniques other than analysis of variance and regression.

Findings also reveal only 31.62% of qualifying articles reported VES (i.e., R^2 and η^2). Although all qualifying articles contained statistical analysis techniques capable of yielding VES, the low rate of variance-accounted-for effect size reporting may indicate that researchers may perceive VES reporting as unimportant. Further, only 8.09% of these articles interpreted VES for practical significance. This finding supports previous observations that researchers may lack the ability to interpret (translate) study results in terms of effect sizes and practical significance (Zakzanis, 2001). The ensuing failure to report VES and interpret them in terms of context-specific practical significance may lead researchers to underreport potentially noteworthy findings (Thompson, 2006; Zakzanis, 2001).

Although this review determined that the majority of published articles using analyses of variance and regression continue to report p values, only a small number report effect sizes. These findings may indicate that health education researchers have not yet begun to acknowledge the limits of statistical significance testing in terms of practical significance and result replicability. An increasing number of health education articles reporting effect sizes would indicate a diffusion of innovation-type effect and facilitate discipline-wide progress (Kirk, 1996). Yet, this trend has not manifested.

IMPLICATIONS FOR PRACTITIONERS

In 2006, an American Educational Research Association (AERA) Task Force issued Standards for Reporting on Empirical Social Science Research in AERA Publications, which mandated effect size reporting and related discussion of practical significance in AERA journals (American Educational Research Association, 2006). This Task Force also encouraged other social science disciplines to adopt these guidelines. Until health education journal editors mandate author guidelines to incorporate effect size reporting, it is recommended that researchers seek out additional information concerning VES reporting and interpretation of practical significance as complements to statistical significance testing (i.e., p values).

Health education researchers should become well-versed in the literature of variance-accounted-for effect size reporting. Increased awareness surrounding this practice will enable researchers to correctly report study findings and promote study replicability. Only when health education researchers find worth in promoting VES will an evolution occur within the body of health education literature. Health education and related social science disciplines will swiftly advance once researchers pursue context-specific, replicable effects (Thompson, 1999a).

REFERENCES

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*(6), 33-40.

American Psychological Association. (1994). *Publication manual of the American Psychological Association*, (4th ed). Washington, DC.

Anderson, D. R., & Burnham, K. P. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management, 64*(4), 912-923.

Buhi, E. R. (2005). The insignificance of "significance" tests: Three recommendations for health education researchers. *American Journal of Health Education, 36*, 109-112.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*, revised edition. New York: Academic Press.

Cohen, J. (1990). Things I have learned "so far." *American Psychologist, 45*(12), 1304-1312.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist, 49*(12), 997-1003.
Elmore, P. B., & Rotou, O. (2001, April). *A primer on basic effect size concepts.* Paper presented at the American Educational Research Association, Seattle, WA.
Fidler, F., Geoff, C., Mark, B., & Neil, T. (2004). Statistical reform in medicine, psychology and ecology. *Journal of Socio-Economics, 33*(5), 615-630.
Goodwin, L. D., & Goodwin, W. L. (1985). Statistical techniques in AERJ articles, 1979-1983: The preparation of graduate students to read educational research literature. *Educational Researcher, 14*(2), 5-11.
Harris, M. J. (1991). Significance tests are not enough: The role of effect-size estimation in theory corroboration. *Theory and Psychology, 1*, 375-382.
Hojat, M., & Xu, G. (2004). A visitor's guide to effect sizes: Statistical significance versus practical "clinical" importance of research findings. *Advances in Health Sciences Education, 9*, 241-249.
Hyde, J. S. (2001). Reporting effect sizes: The roles of editors, textbook authors, and publication manuals. *Educational & Psychological Measurement, 61*(2), 225-228.
Keselman, H. J., & Huberty, C. J. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA and ANCOVA. *Review of Educational Research, 68*(3), 350-386.
Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational & Psychological Measurement, 56*(5), 746-759.
Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational & Psychological Measurement, 61*(2), 213-218.
McLean, J. E., & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *Research in the Schools, 5*(2), 15-22.
Mitchell, R. E., Ash, S. L., & McClelland, J. W. (2006). Nutrition education among low-income older adults: A randomized intervention trial in congregate nutrition sites. *Health Education & Behavior, 33*(3), 374-392.
Nix, T. W., & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the Schools, 5*(2), 3-14.
Tapia, M., & Marsh, G. E., II. (2002). Interpreting and reporting effect sizes in research investigations.
Thompson, B. (1998). *Why "encouraging" effect size reporting isn't working: The etiology of researcher resistance to changing practices.*
Thompson, B. (1999a). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology, 9*(2), 165-181.
Thompson, B. (1999b). Statistical significance tests, effect size reporting and the vain pursuit of pseudo-objectivity. *Theory & Psychology, 9*(2), 191-196.
Thompson, B. (2002a). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling & Development, 80*, 64-71.
Thompson, B. (2002b). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 25-32.
Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach.* New York: Guilford.
Thompson, B. (Ed.). (2003). *Score reliability: Contemporary thinking on reliability issues.* Newbury Park, CA: Sage.
Tolma, E. L., Reininger, B. M., Evans, A., & Ureda, J. (2006). Examining the Theory of Planned Behavior and the construct of self-efficacy to predict mammography intention. *Health Education & Behavior, 33*(2), 233-251.
Vacha-Haase, T. (2001). Statistical significance should not be considered one of life's guarantees: Effect sizes are needed. *Educational & Psychological Measurement, 61*(2), 219-224.
Vacha-Haase, T., & Thompson, B. (1998). Further comments on statistical significance tests. *Measurement & Evaluation in Counseling & Development, 31*(1), 63-67.
Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology, 51*(4), 473-481.
Watkins, D. C., Rivers, D., Rowell, K. L., Green, B. L., & Rivers, B. (2006). A closer look at effect sizes and their relevance to health education. *American Journal of Health Education, 37*(2), 103-108.
Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594-604.
Zakzanis, K.K. (2001). Statistics to tell the truth, the whole truth, and nothing but the truth: Formulae, illustrative numerical examples, and heuristic interpretation of effect size analyses for neuropsychological researchers. *Archives of Clinical Neuropsychology, 16*, 653-667.